

A comparison of sample survey measures of earnings of English graduates with administrative data

JACK BRITTON

Institute for Fiscal Studies, London

NEIL SHEPHARD

Department for Economics and Department of Statistics, Harvard University

ANNA VIGNOLES

Department of Education, University of Cambridge

March 2, 2018

Abstract

Administrative data sets are increasingly used in research due to their excellent coverage and large scale. However, in the UK the use of administrative data on individuals' earnings, and particularly graduates' earnings, is novel. Understanding the strengths and weaknesses of such data is important as they are set to be used extensively for research purposes and to inform policy. Here we compare survey based labour earnings data from the UK's Labour Force Survey (LFS) with UK government administrative sources of individual level earnings data, focusing separately on young (up to age 32) graduates and non-graduates. This type of administrative dataset has few sample selection issues, is longitudinal and its large samples mean the earnings of sub-populations can potentially be studied with low error. Overall we find a similar share of individuals with zero earnings in the LFS and administrative data, but a considerably higher share (conditional on working) earning below £8,000 in the administrative data. The LFS has generally higher earnings right through the distribution, though above the median a large share of the differences can potentially be explained by employee pension contributions. We also find considerably larger gender difference in the survey data. The findings hold for both graduates and non-graduates. These differences are substantively important and suggest different conclusions about the gender wage gap, the graduate earnings premium and the extent of earnings inequality.

Keywords: Administrative data; Graduate earnings; Labour force survey; Student loans.

1 Introduction

A rich literature has shown the power of administrative tax records to better understand the earnings of subpopulations (e.g. Chetty et al. (2014a,b)). Such data have comprehensive coverage, clearly defined income categories and individual (or household) level data that stretches over significant periods of time. Given these advantages, as discussed in Savage and Burrows (2009), Webber (2009) and Card et al. (2010), there is a growing literature on the application of large scale administrative data to understand the outcomes from education (see Figlio et al. (2015), Black et al. (2005), Bhuller et al. (2017) and Carneiro et al. (2013) for illustrations of the use of such data). However, whilst administrative data have been used to good effect to study labour markets

in a number of different countries, their use in the UK is in their relative infancy and there has been little work establishing the quality of such data.

Here we build and document a new database we call the “Golden Sample” (GS) that links administrative tax records for young (up to age 32) individuals to their Student Loan Company (SLC) records. This enables us to investigate the earnings of English graduates. We compare the GS’s summary statistics with corresponding results from a well established government funded labour market sample survey, the Labour Force Survey (LFS), exploring the relative strengths and weaknesses of both of these sources of data. Such data are set to take a more prominent role in UK policy making in years to come. For example, current estimates of the long-run costs of income-contingent student loans in the UK, which require the forecasting of graduate earnings several years into the future, are largely based on survey data, and the LFS in particular. The administrative dataset that we introduce here is set to be used by the UK Government to investigate these long-run costs, as it provides rich earnings information with long panels with large sample sizes and links to higher education providers and subject choice that allow detailed breakdowns of the cost of loans by sub-populations. Documenting the differences between, and relative advantages of, the administrative and survey data, particularly for graduates’ earnings, is therefore of great importance for researchers and policy makers. For comparison purposes, we also build a less rich dataset of UK based non-graduates, which we call the “Silver Sample” (SS), and compare it to non-graduate LFS data.

We compare our administrative data to the LFS, a survey commonly used to estimate graduate earnings (e.g. Walker and Zhu (2011)) and other labour market measures (e.g. Cribb and Joyce (2015)). Overall we find a similar share of individuals with zero earnings in the LFS and the administrative data, but a considerably higher share (conditional on working) earning below £8,000. The LFS has generally higher earnings right through the distribution, though above the median a large share of the differences can potentially be explained by employee pension contributions. These findings are robust to whether we are looking at graduates or non-graduates.

These differences have implications for future research and public policy. The administrative datasets are the official earnings records for an individual and hence the earnings that are relevant for the loan repayment calculations and for tax receipts. Further, we also believe that the administrative data are more reliable than the LFS, at least conditional on earnings being greater than £8,000. However, we find that the high share of individuals earning between £0 and £8,000 and in particular the lack of gender differences in the lower part of the distribution in the administrative data is not just inconsistent with the LFS but also the Family Resources Survey (FRS), another commonly used survey for studying earnings. There are a number of potential explana-

tions for this, including: sample selection or non-response bias that results in low earners being under-represented; measurement error, in particular annualising earnings from sometimes shorter periods and the treatment of those with variable income; and the inclusion of sources of income other than earned income, such as salary sacrifice pension contributions. Alternatively, the administrative data may suffer from under-reporting of income in order to avoid paying tax or to qualify for benefits, and also has issues caused by the inclusion of all English-domiciled borrowers rather than graduates. All these issues are discussed in more detail below, when we use the data to measure the gender wage gap, the graduate wage premium, and earnings inequality amongst graduates and non-graduates. We show how important conclusions made about the economic advantages or otherwise of taking a degree differ, depending on which data source is used. This paper serves to improve our understanding of the different features of the two data sources and, whilst we conclude that the administrative data does indeed have considerable advantages over the survey data, we also highlight limitations that need to be understood by researchers if such data are to be used to best effect.

The rest of the paper is laid out as follows. In Section 2 we review the literature and in Section 3 we detail the linkage of administrative data. In Section 4 we discuss the UK Labour Force Survey and summarise the key differences between data sources. Section 5 compares earnings distributions of LFS graduates versus the GS, LFS non-graduates versus the “non-HE” sample (a corrected version of the SS) and the overall LFS population versus the combined GS and SS populations. Section 6 makes applied comparisons, investigating differences in the gender wage gap, the graduates to non-graduates earnings ratio and earnings inequality, and Section 7 concludes. An Online Appendix contains various additional results referred to in this paper.

2 Literature

This paper builds on a significant literature (e.g. Bound et al. (2001a), Abowd and Stinson (2013), Koijen et al. (2015)) which has discussed the problems of using sample surveys, particularly in relation to measuring income, comparing their results to some administrative data. There are also specific data collection issues in the LFS. Skinner et al. (2002), for example, have found significant discrepancies in earnings estimates for the low paid using different ways to calculate hourly pay in the LFS. Traditionally, hourly pay in the LFS was calculated using weekly pay received divided by usual working hours. More recently, the LFS have included an hourly rate of pay variable. They conclude that the latter has less measurement error but is missing for a significant proportion of the sample. Imputing earnings for those with missing data on the hourly rate of pay variable leads to substantially reduced estimates of the proportion who are low paid. Bound et al. (2001b)

discuss sources of error in earnings surveys, concluding that self-reported annual earnings tend to have less error than disaggregated measures, such as hourly or weekly earnings (see also Duncan and Hill (1985)). Bound et al. (2001b) also found evidence that survey errors are mean-reverting. There was mixed evidence on whether graduates or individuals with more human capital were more likely to report their earnings with error, though some studies that have compared survey measures with administrative records have found a positive correlation between true earnings and error in earnings (e.g. Rodgers et al. (1993)). Bound et al. (2001b) found limited evidence that respondents with very high earnings tend to under-report their earnings and those with very low earnings over-report theirs, rejecting the theory of “social desirability bias” where individuals report with bias so as to appear less different. However, they did find non-negligible measurement error in measures of schooling and highest education level: individuals can misreport or misremember their years of schooling or highest level of qualification. This measurement error in reported schooling levels will in turn cause measurement error in estimates of earnings differences by education level, even if individuals report their earnings correctly.

A review by Moore et al. (2000) that focused on sources of error in earnings measures in official surveys suggested a wide range of different sources of bias. Non-response is an issue. Respondents may not completely understand the different definitions of earnings being used (e.g. in the LFS they are asked for earnings both “before deductions” and net pay “after deductions”). Questions may not be precise about the inclusion of pension contributions and childcare allowances, and individuals may have recall problems depending on the period being asked about. Whilst it is well known that earnings data collected with a single question are subject to extensive measurement error (e.g. Micklewright and Schnepf (2010)), even when more complex survey designs are used it remains a challenge to design high quality instruments for measuring earnings in surveys, particularly if household members are being asked to report on the earnings of others (this is the case with the LFS: in practice we actually find removing proxy responses makes little difference to our conclusions). Indeed previous work has identified differences in earnings estimates across a number of survey based sources. For example, comparing UK individual survey data on earned income (from the Family Expenditure Survey and the General Household Panel Survey) with surveys of businesses (Annual Survey of Hours and Earnings, which is based on a 1% sample of employee jobs taken from HM Revenue and Customs (HMRC) Pay As You Earn (PAYE) records, with information on earnings and hours obtained from employers) have tended to find that the former under estimate the earnings of respondents as compared to the latter (Atkinson et al. (1981, 1982), Devereux and Hart (2010)). It should be noted that although ASHE does sample from HMRC tax records, it is based on a survey of employers only, so does not include self-assessed earnings nor those not paid

in the reference week. It is also still based on a survey methodology and suffers from non-response, meaning it is not directly comparable with our work.

This paper contributes to this literature in several ways. First, the paper compares the distribution of earnings in the administrative data with the LFS and highlights the potential limitations of both datasets, which is an issue of increasing relevance as administrative data sources start to become more readily available for policy makers. Second, it provides evidence on the level and variation of UK graduate earnings using this new high quality data source (Naylor et al. (2016), Walker and Zhu (2011)). Third, it highlights the rich potential of this dataset for understanding inequality in earnings, adding to the large body of work on this issue (Cunha and Heckman (2016) provide a comprehensive summary). Fourth, it provides UK evidence on the gender wage gap particularly amongst graduates, again building on previous UK empirical work which has often relied on survey data and the LFS specifically (Machin and Puhani (2003), Chevalier (2007)).

3 Our administrative databases

3.1 The Golden Sample (GS)

The GS is a database we built, using National Insurance Numbers (NINOs) to hard link three datasets: data from the SLC and Pay As You Earn (PAYE) and Self-Assessment (SA) databases from Her Majesty’s Revenue and Customs (HMRC). This provides us with a large longitudinal database on UK earnings for individuals domiciled in England upon application to HE, who received loans from the SLC.

The two HMRC datasets arise because the UK has two types of income tax forms. The significant majority of tax payers use the PAYE system, which is operated by employers who withhold income and other employment taxes and report the earnings and deductions made to HMRC. This means the majority of UK citizens do not themselves file tax forms; Pope and Roantree (2014) report that around 90% of UK income tax is collected through the PAYE system. For those with more complicated tax affairs (e.g. high incomes, self-employed, owning a business, having significant investment accounts, being in a professional partnership) HMRC requires them to file a set of SA forms. Individual taxpayers can also opt to submit SA forms.

The UK runs an individual tax filing system with no option to file as a household. Thus UK administrative data will be good for studying individuals’ earnings but, unlike the US (e.g. Guvenen et al. (2014)), not good for household earnings. HMRC does have address information which would allow the fuzzy linkage into households, but we do not have access to this information. We therefore focus on individual rather than household earnings.

3.1.1 Earnings data

Our focus is on earned labour income, so we defined this as the sum of employment income, profits from partnerships and profits from self-employment declared to HMRC. Clearly some aspects of the returns from a partnership are due to the capital risk a partner is exposed to, but we cannot break that component out here and so take profits from partnerships as earnings.

The SA databases also contain information on trust income, profits on share transactions, profits from land and property, UK dividends, pension income, life policy gains, “other” income, bank and building society interest and total income, all of which we exclude from earned income as they measure non-employment income. We wanted to include foreign income from employment and savings, but the calculation involved various delicate deductions, so we excluded it.

We do not make a record of any deductions tax payers make, e.g. capital losses on investments, nor of any tax free allowances individuals may have. We also do not account for employers’ and employees’ tax free pension contributions as labour earnings as UK tax forms only record pension income and not pension contributions.

When we have both PAYE and SA earnings we use the SA data, as HMRC regard the SA records as definitive (noting that a SA form will include PAYE income). If an individual has no reported earnings then we take their earnings as zero. This is likely to miss some earnings for very low earners who do not have to return a PAYE form and who may not be asked to complete a SA form (although note that they have a legal responsibility to report this income). All earnings are converted into October 2012 prices using the Consumer Price Index (CPI).

3.1.2 Student Loan Company (SLC) data

The SLC has offered income contingent loans to all UK domiciled HE students since 1998. The take-up rate amongst eligible students during this period is around 85-90% overall, a rate that has remained relatively stable (author’s own calculations based on overall students numbers from the SLC “Student Support for higher education in England” archived series). Not all individuals receiving a loan from the SLC will be studying for first degrees, as individuals can access loans for foundation degrees, Higher National Diplomas (HNDs) and lower undergraduate qualifications. The dataset we received from SLC does not have any indicators to split individuals into these different groups. We observe the final degree for which an individual qualifies for a loan. So, for example, for someone attending a HE institution for a term before dropping out and re-starting at a different institution sometime in the future, only their second degree is observed so long as they borrowed again (though the date they started in HE is the first degree start date).

The dataset only includes individuals who borrowed from the English part of the SLC - meaning

they were domiciled in England upon application - between 1998 and 2010 and covers around 2.6M former borrowers who are qualified to be in repayment, which happens in April of the year after they leave HE. We have no data on those who are still in HE and have insufficient earnings to qualify them for repayment, which results in a decline in our cohort sizes for more recent student cohorts (see Table 1). Note that we only observe borrowers and not whether individuals graduate, resulting in individuals who borrow from the SLC but subsequently drop out being inaccurately defined as graduates (throughout, we use the terms “borrowers” and “graduates” interchangeably). During this period the drop out rate from UK universities for those who enroll was around one in ten, including mature entrants (taken from HESA performance indicators data series, where HESA measures drop out by those who attended for at least 90 days before dropping out).

3.1.3 Linking the administrative datasets

Primarily due to computational limitations, HMRC have allowed us to link 10% of individuals in the SLC data to the tax data, with the 10% selected based on two digits within each individuals’ randomly allocated National Insurance Number (NINO). HMRC use the same 10% for much of their own analysis. Because we have the full sample of borrowers, our 10% sample includes the small fraction of individuals who never file with HMRC.

We call this 10% matched sample the GS. We have up to eleven tax years (note that the tax year runs from April 6th-April 5th each year) in the dataset for each individual, from 2002/03 - 2012/13, although the majority of our focus here is on the 1998-2003 cohorts in 2008/09-2012/13 in order to give individuals sufficient time to complete their degrees and enter the labour market after starting their HE course. Once submitted to the HMRC, UK tax forms are highly confidential and access to them is restricted by Parliamentary statutes. We have been given access to an anonymized version of the data and our work was carried out in a highly secure data enclave within a HMRC facility. All outputs are checked by officials to ensure they cannot be disclosive of any individual’s information.

The matching is a hard link based on the individuals’ NINO, which is available in both datasets and the quality of which is checked many times (for more detailed information on this see Britton et al. (2015)). These data therefore do not suffer from the weaknesses of some other linked administrative data sets; for example Chetty et al. (2014a) report linkage rates close to 90% using fuzzy matching, based on date of birth, state of birth, names and gender, between school reports and tax records and just under 98% for matching parents to children (although it should be noted the Chetty et al. (2014a) do have considerably larger sample sizes).

A drawback is that when former students become non-resident for UK tax purposes, HMRC may lose contact with them and generally will only record earnings from UK sources as these are

their UK taxable earnings. We will express the earnings of such students as 0 in our reports if HMRC records it as 0, which clearly may underestimate their true earnings.

3.1.4 Basic summaries of the Golden sample

The GS has 263,052 members, covering cohorts from 1998 to 2011. We focus on the 2008-09 to 2012-13 tax years. It should be noted that this was a financially difficult period. The GS is detailed for 2011/12 in Table 1. There are around 24,000 students in each cohort, with the smaller 1998 figure reflecting slow uptake of the new income contingent student loans and the decline at the end reflecting the fact that individuals have not entered repayment (i.e. left HE) by 2011/12. The student numbers align with HESA statistics for 2007/08, which state that around 325,000 UK domiciled students were studying in England. Our 10% sample is 25,000 students in this year, meaning a cohort size of around 250,000 borrowers. Around 15% of the English students do not borrow (taking us to 295,000), while the remaining students would be non-English UK students studying in England.

Each individual potentially has a SA and a PAYE tax record in each tax year, but may have neither. By construction, we are able to state that if they have neither a SA nor a PAYE record then they have no UK tax return at all - note that unlike the US, in the UK it is not legally necessary to file a tax form if your income is indeed zero, although it is required for any amount above 0. We will record such non-filers as having zero earnings. We end up with the GS for whom we have earnings data from the PAYE database, the SA database or both.

Cohort	All				Male				Female			
	Golden	PAYE	SA	Either	Golden	PAYE	SA	Either	Golden	PAYE	SA	Either
1998	14,487	11,646	2,310	12,226	6,927	5,528	1,351	5,875	7,560	6,118	959	6,351
1999	22,621	18,410	3,447	19,354	10,590	8,529	1,912	9,063	12,031	9,881	1,535	10,291
2000	23,506	19,214	3,425	20,176	10,853	8,761	1,908	9,322	12,653	10,453	1,517	10,854
2001	23,924	19,921	3,108	20,818	11,025	9,060	1,759	9,625	12,899	10,861	1,349	11,193
2002	23,891	20,104	2,814	20,906	11,060	9,156	1,576	9,642	12,831	10,948	1,238	11,264
2003	23,972	20,387	2,447	21,097	11,024	9,315	1,314	9,726	12,948	11,072	1,133	11,371
2004	23,577	20,367	2,266	20,997	10,767	9,163	1,251	9,526	12,810	11,204	1,015	11,471
2005	25,103	21,800	2,085	22,397	11,439	9,822	1,141	10,183	13,664	11,978	944	12,214
2006	25,383	22,149	1,864	22,589	11,340	9,749	992	10,024	14,043	12,400	872	12,565
2007	25,352	22,303	1,527	22,694	11,292	9,746	774	9,981	14,060	12,557	753	12,713
2008	20,847	18,154	1,039	18,430	8,990	7,704	531	7,872	11,857	10,450	508	10,558
2009	6,510	5,386	426	5,485	3,029	2,452	215	2,509	3,481	2,934	211	2,976
2010	2,993	2,477	152	2,511	1,334	1,082	72	1,101	1,659	1,395	80	1,410
2011	851	721		724	360	291		294	491	430		430
All	263k	223k	27k	230k	120k	100k	15k	105k	143k	123k	12k	126k

Table 1: Number of Golden sample (10% sample of loan database) borrowers and tax data in 2011-12. PAYE (Pay As You Earn) and SA (self-assessment) denotes databases. Either denotes being in either PAYE or SA or both. Cohort denotes the first year the borrower received a loan from the SLC.

Table 1 gives the break down of different types of tax forms for 2011/12 by gender and cohort. It shows a significant majority of borrowers are female for all cohorts, reflecting greater HE partici-

pation amongst women. In more recent cohorts there is very little SA data since it is higher earners and the self-employed that are more likely to use SA, both of which become more likely with age. There are some people, mostly self-employed, who appear only in the SA data (e.g. in 1998 of the 14,487 individuals in the GS, 12,226 have tax records for that year and 1,730 had only SA records - this is equal to $11,646 + 2,310 - 12,226$) and a considerably higher rate of SA for males.

Table 2 shows the percentage of individuals who file no tax form at all during 2011/12, and the share with no and low earnings, by cohort (with the median age of the cohort indicated). The columns are cumulative, so the share with earnings $< £8,000$ includes those with zero earnings and those with no filed tax form. For those with no form, we assume the individual has zero taxable income in the UK. The rate of not filing initially decreases moving up through the cohorts, but then increases. There is little gender difference in the not filing rate, even as the cohort reaches their early thirties, which is surprising given evidence on the unequal split of childcare responsibilities - one possible explanation might be predominantly female individuals paying National Insurance contributions even when they have zero earnings so as to preserve pension benefits in later life, although we do not attempt to quantify this here. There is a sizeable group of people in the databases with returns of zero income (given by subtracting the share with no form from the share with earnings of zero - for example for the 1998 cohort, 2.7% of women file returns of zero earnings). This might arise, for example, from employers filing PAYE returns for former employees. Again there is very little difference by gender.

Median age	Cohort	% No tax form			% Earnings = £0 (or no form)			% Earnings $< £8,000$ (includes 0s & missings)		
		All	Male	Female	All	Male	Female	All	Male	Female
31	1998	13.0	12.6	13.3	15.6	15.2	16.0	27.3	26.7	27.9
30	1999	11.7	11.4	11.9	14.4	14.4	14.5	26.2	25.7	26.7
29	2000	11.4	11.2	11.5	14.2	14.1	14.2	26.1	25.7	26.5
28	2001	10.1	9.9	10.3	13.0	12.7	13.2	25.0	24.5	25.5
27	2002	9.6	9.9	9.3	12.5	12.8	12.2	25.3	25.5	25.0
26	2003	9.0	8.9	9.0	12.0	11.8	12.2	25.8	25.4	26.1
25	2004	8.0	8.3	7.7	10.9	11.5	10.5	25.9	26.8	25.2
24	2005	7.5	7.4	7.5	10.8	11.0	10.6	29.1	30.3	28.2
23	2006	7.5	7.8	7.2	11.0	11.6	10.5	34.3	36.3	32.6
22	2007	7.0	7.8	6.3	10.5	11.6	9.6	43.2	45.1	41.8
21	2008	8.4	9.1	7.8	11.6	12.4	11.0	61.6	63.2	60.4
21	2009	10.9	11.6	10.4	15.8	17.2	14.5	61.1	64.6	58.0
20	2010	11.0	12.0	10.2	16.1	17.5	15.0	67.9	72.0	64.6
18	2011	10.1	13.1	7.9	14.9	18.3	12.4	90.6	90.6	90.6

Table 2: Golden Sample for 2011-12. % of individuals with no filed income form & the % with no or low earnings. Columns are cumulative so the share with earnings $< £8,000$ includes those with earnings = £0 and those with no form. Median age does not decrease by one each year in the GS because of small sample sizes and variation in the ages of HE leavers (since individuals only enter our dataset once they have left HE).

The rate of borrowers with zero earnings appears to be high, accounting for over 14% of individuals aged around 30. However, this figure is comparable to SLC official statistics (which are

not perfectly equivalent, as they include EU borrowers). These show that of the 2001 cohort in 2013/14 (as close to the equivalent for the 1999 cohort in 2011/12 as we could get), 9% still have debt but have no employment. Approximately 1.4% of individuals have had their loans written off due to death, disability or bankruptcy and 37% have cleared their debts. Some individuals in each of the latter groups will have zero earnings but would not be incorporated in the 9% figure. If all of those with debt written off due to death disability or bankruptcy were on zero earnings, that would be 10.4% of borrowers, and if just 7% of those with cleared debts were also on zero earnings, that would be take us to around 13%.

The remaining difference can most likely be explained by individuals moving abroad. Table 3 summarises some additional SLC information on this. This shows that around 1% of the 1999 cohort were abroad and in repayment in 2011/12. These data are incomplete, as the SLC does not continue to track individuals country of residence once they are out of repayment - we therefore also show figures for individuals who have been abroad at any point (which includes those currently abroad). This is around 4% for the 1999 cohort in 2011/12. The table also shows the share of individuals with no and low (less than £8,000) earnings. Around 80% of those currently abroad have earnings in the UK below £8,000, while more than half of those ever abroad do. This shows that some individuals still file while they are abroad, but it also suggests that more than 1% of individuals are abroad at any given time. Combining this with the 13% figure above, this therefore gets us close to SLC official records. The UK Department for Education has also started to separately use HMRC administrative data on earnings (with some notable differences; they do not yet use SA data, and they are not able to hard link datasets to identify graduates using NINO) and their calculations suggest a similar proportion of graduates with zero earnings.

Cohort	% abroad	% been abroad	Of those abroad		Of those been abroad	
			Earnings= £0	Earnings<£8,000	Earnings=£0	Earnings<£8,000
1998	1.0	4.4	73.2		40.8	50.6
1999	1.1	4.2	69.8	86.8	42.7	52.2
2000	1.2	4.0	61.7	80.7	42.0	53.9
2001	1.2	4.1	66.2	78.5	42.5	52.2
2002	1.3	3.8	58.4	78.1	39.5	52.1
2003	1.4	3.7	57.0	73.8	40.8	55.1
2004	1.4	3.8	47.7	71.8	36.4	54.9
2005	1.4	3.4	55.9	82.6	39.3	61.9
2006	1.4	2.6	51.2	85.3	43.1	73.5
2007	1.4	1.9	43.8	86.0	44.0	82.9
2008	0.6	0.7	34.1		35.3	

Table 3: SLC in repayment and living abroad data in 2011/12. Abroad is an indicator for being overseas and in repayment according to SLC records. Been abroad is an indicator for abroad and in repayment *or* have been in this state at some point in the past. Figures are excluded where implied sample sizes are too small.

Table 2 also records the percentage of borrowers with incomes below £8,000. This level was selected since it is approximately the level of earnings at which individuals start to pay National

Insurance Contributions and income tax (Pope and Roantree (2014)), meaning the administrative data are more likely to be reliable above this level. Around a quarter of borrowers earn less than £8,000 around their late twenties and early thirties, with again only a relatively small difference between genders. This finding is stark, and we return to it below.

One concern is under-reporting of earnings, an issue that might be a particular problem for the self-employed, for whom it is easier to move income into other forms as there is no employer based filing which can be used to independently verify the income. Indeed Her Majesty’s Revenue and Customs (2014) have estimated the amount of uncollected tax caused by the under-reporting of income, finding a tax gap of around 17% for self-assessed taxes (with around 25% of SA taxpayers under-reporting their earnings) and 1.5% for PAYE taxes. Since the vast majority of our data comes from PAYE sources, and the majority of those with SA reports also have most of their earnings recorded through employer based PAYE records (i.e. the “P60” form), the main vulnerability of the tax data is therefore to under-reporting from the fully or partially self-employed.

Table 4 quantifies the degree of self-employment in this dataset, showing how it varies with cohort and gender. Around 10% of our sample are either fully or partially self employed. We have not made any correction to the raw HMRC data in our analysis to take this under-reporting into account, though we would obviously expect this to bias our estimates for this group downwards. The proportion of borrowers who only have earnings from self-employment is roughly 1-3%, clearly increasing with age and with a higher rate for men than women. Of these, around 35% of men report labour earnings of below £8,000, while the equivalent figure for women is almost 60%. A higher rate of partial self-employment is recorded, again with males having higher incidence than females. Amongst these individuals, women again have a considerably higher chance of having low earnings. We refer back to this when considering the high incidence of low earnings in the tax data in Section 5.

3.2 The Silver Sample (SS)

The HMRC and SLC linking also yields a sample of people who did not take out loans from the English part of the SLC. The significant majority of these UK people are non-graduates. This database is called the “Silver Sample” (SS).

The SS is built by taking the 10% NINO sample (which, as described above, is a random 10% sample of the population) in the tax data and removing all those who appear in the SLC database. Specifically, this means the SS consists of anybody who appears at any point in the PAYE or SA tax data between 2008/09 and 2012/13 inclusive, is in the 10% NINO sample, and does not appear in the SLC dataset (meaning they did not borrow to go to university). For each person in this population we know their age, gender and earnings (including type of earnings) only. Then for

Median age	Cohort	Only partly self-employed						Entirely self-employed					
		Of all (%)			Of SE part: % earnings < £8,000			Of all (%)			Of SE only: % earnings < £8,000		
		All	M	F	All	M	F	All	M	F	All	M	F
31	1998	6.4	7.1	5.7	33.4	27.1	40.7	3.6	4.4	2.8	44.9	35.4	58.8
30	1999	6.5	7.3	5.8	34.6	30.3	39.3	3.8	4.5	3.1	46.4	39.4	55.3
29	2000	6.6	7.5	5.8	33.8	31.7	36.1	3.7	4.6	2.9	46.7	42.9	51.9
28	2001	6.2	7.5	5.1	34.3	31.7	37.5	3.5	4.7	2.5	47.6	43.4	54.4
27	2002	5.8	6.9	5.0	35.9	35.5	36.3	3.3	4.3	2.4	47.2	46.3	48.7
26	2003	5.4	6.1	4.8	37.9	33.9	42.1	3.0	3.6	2.5	52.1	46.6	58.9
25	2004	5.2	6.2	4.3	38.8	36.2	41.9	2.8	3.6	2.1	51.7	47.7	57.6
24	2005	4.9	5.9	4.1	41.3	41.5	41.1	2.6	3.3	2.0	58.3	55.3	62.6
23	2006	4.3	5.1	3.7	47.6	46.6	48.8	2.2	3.0	1.5	63.1	59.7	68.5
22	2007	3.8	4.4	3.4	54.7	50.7	58.8	1.9	2.5	1.5	68.2	61.2	77.7
21	2008	3.1	3.6	2.7	67.9	68.3	67.5	1.7	2.4	1.2	78.1	78.0	78.1
21	2009	3.4	4.0	3.0	62.5	63.3	61.5	1.8	2.2	1.4	85.2	86.4	83.7
20	2010	2.8	3.4	2.4	61.9			1.3			87.5		

Table 4: Golden Sample self-employment: cohort who are only partially self-employed (not those fully self-employed) and those entirely self-employed. Also given are the corresponding % who have low earnings. Earnings is all earnings from work, not just from the self-employed part. Results are for the 2011-12 tax year. See footer to Table 2 to explain pattern for median age.

each cohort and gender we have sampled this new population to produce a database with the same age profile as the SLC database. This results in a large database which the HMRC systems have difficulty coping with. We therefore randomly select a subset of the SS, keeping approximately two members of the SS for every one in the GS, which roughly halved the overall size of the SS.

Median age	Cohort	% No tax form			% Earnings = £0 (or no form)			% Earnings < £8,000 (includes zeros & missings)		
		All	Male	Female	All	Male	Female	All	Male	Female
31	1998	22.1	21.5	23.0	27.3	26.7	27.9	46.3	43.3	49.9
30	1999	22.6	21.3	24.2	27.7	26.6	29.0	47.5	43.8	51.9
29	2000	23.5	21.8	25.5	28.5	27.0	30.4	48.8	45.2	53.2
28	2001	24.3	22.4	26.5	29.1	27.6	31.0	49.7	46.1	54.0
27	2002	24.8	23.1	26.8	29.7	28.3	31.4	51.2	47.9	55.1
26	2003	25.0	23.2	27.2	29.9	28.2	31.9	51.9	48.5	55.8
25	2004	24.9	22.7	27.5	30.1	28.1	32.5	52.9	49.8	56.6
24	2005	24.2	21.8	27.0	29.3	27.3	31.7	53.8	51.2	56.9
23	2006	23.7	21.4	26.4	29.0	26.9	31.4	55.8	53.4	58.6
22	2007	22.8	20.3	25.6	28.2	25.9	30.9	58.6	55.7	61.9
21	2008	21.6	19.4	24.1	27.8	25.4	30.5	61.6	59.0	64.5
21	2009	20.4	19.5	21.3	26.4	25.6	27.3	64.2	62.0	66.7
20	2010	18.4	17.1	19.9	24.4	23.1	25.8	68.8	66.0	71.8

Table 5: Silver Sample database for 2011-12. % with no filed income tax form and % with no and low earnings. Median age does not decrease by one each year in the SS because the age distribution is matched exactly to the GS (see footer to Table 2).

Summaries of the characteristics of the SS are given in Table A1 in the Online Appendix A. There are more men in the SS, reflecting the fact that there are more women in the GS, and the rate of SA is lower in the SS than the in GS (e.g. in 1999 the GS SA rate is about 15%, while for the SS it is about 11%). Table 5 shows the rate of low pay in the SS is roughly twice as high as for

the GS, with 45% of non-graduates with earnings below £8,000, compared with 25% of graduates. There is also more of a gender difference than in the GS, with around 50% of females in their early thirties earning below £8,000 in the SS, compared with 43% of males.

3.2.1 Correcting the Silver Sample

There are three major issues with the SS (that do not apply to the GS). First, it misses people who have no tax record at all in either of the PAYE or the SA datasets from 2008/09 to 2012/13. Second, immigrants entering the country and being assigned a NINO will be included and are (at least in principle) also included in the LFS. A problem is created if the individual is not in the country for the entire five year period. For example, an individual who enters the country in 2012/13 would be recorded as having no tax form and hence zero earnings in each of the other years. Third, the SS includes graduates from England who did not borrow (around 15% of English graduates) as well as graduates (and non-graduates) from Scotland, Wales and Northern Ireland.

For the first two of these issues, there is little that can be done, since Table 2 shows that a high share of graduates have no tax form, so omitting everybody with no form would dramatically underestimate the share with no earnings. Consequently, we focus much of our later analysis on those with positive earnings only. This resolves the issue with no form and considerably reduces the problem with immigration (it does not completely remove it, however, as immigrants may file a form despite only being present in the country for a fraction of the tax year).

We can correct the third issue by effectively re-weighting the earnings distribution of the SS based on the share of graduates we believe are present in the sample. The result is called the “non-HE” sample. Let $F_S(y) = \Pr(Y \leq y)$ be the distribution function of SS earnings Y for a specific cohort and gender. F_{HE} will be the corresponding result for the subset that went into HE and F_{nonHE} is the result for the others. We write ω as the proportion of graduates in the Silver sample, then by construction.

$$F_S(y) = \omega F_{HE}(y) + (1 - \omega) F_{nonHE}(y), \quad \omega \in [0, 1].$$

We now make the assumption that $y \in R_{\geq 0}$, then $F_{HE}(y) = F_G(y)$ where F_G is the distribution function from the GS. This says that the distribution of earnings of the graduates in the SS matches the distribution of earnings in the GS - that is the GS well represents all graduates, not just English borrowers. It is important to note that F_G is likely to underestimate earnings for English graduates who do not borrow as we might expect this group to come from wealthy families and to subsequently have higher than average earnings themselves, but it is difficult to quantify this underestimation (especially as there are other reasons people may not borrow that might not be positively correlated with subsequent earnings, such as debt aversion).

Under these assumptions, for $y \in R_{\geq 0}$,

$$F_{nonHE}(y) = \frac{F_S(y) - \omega F_G(y)}{(1 - \omega)}, \quad E_{nonHE}(Y) = \frac{E_S(Y) - \omega E_G(Y)}{(1 - \omega)}, \quad f_{nonHE}(y) = \frac{f_S(y) - \omega f_G(y)}{(1 - \omega)}.$$

Since we can estimate F_S and F_G from the data, we simply need ω in order to make this correction. Using a combination of data from the Office for National Statistics (ONS), government records and SLC data we estimate that ω is equal to around 0.14 for men and 0.21 for women (see the Online Appendix B).

Around a half of these are non-borrowers from England and the rest are all of the graduates from Scotland, Wales and Northern Ireland. Hence the SS will typically overestimate the distribution of earnings for non-graduates, as HE graduates are much more likely to be very high earners than non-HE people, yielding a large bias if we use it to learn about the upper tail or mean of earnings for non-graduates. However, at the centre of the distribution and in the left hand tail it is likely to be a good approximation. It should be more accurate for men than for women because the estimated share of graduates in the SS is lower for men than for women.

4 The Labour Force Survey (LFS)

The LFS has a rolling five wave design, with 20% of the overall sample replaced with new respondents each quarter. Individuals are surveyed for five consecutive quarters, meaning five waves of data may be available for one person with the first and fifth waves one year apart. Earnings questions only appear in waves 1 and 5. Many people will take the survey but, as is often the case, not provide information on earnings while answering other questions.

Table 6 shows the sample sizes for the LFS between April 2011 and March 2012 (i.e. quarters 2-4 from 2011 and quarter 1 from 2012) for different ages which are the closest match to SLC cohort, by gender and graduate status. The latter is determined from the “highest qualification” question in the LFS, which we choose to align as closely to the HMRC data as possible, meaning we include all courses that are eligible for student loans (our code is available on request; in practice our definition also aligns closely with Walker and Zhu (2011)). Due to the lack of more information on higher education, cohorts are assigned to individuals based on their age on August 31st in a given year, which we observe in the LFS special license access dataset. For example, individuals who were 18 on September 1st 1998 are assigned to the 1998 cohort. We focus on individuals who are domiciled in England at the point of survey (rather than on the point of application to HE) and we have included proxy earnings responses, where individuals complete the earnings questions on behalf of a family member (although previous work (Wilkinson (1998)) has suggested proxy earnings might underestimate true earnings, our findings are not sensitive to this decision).

The Table gives the raw, unweighted sample sizes, which means they are affected by non-response. This drives the larger number of women and the lower share of graduates in the data. We deal with non-response using LFS population weights in our subsequent comparisons.

Cohort	Graduates				Non-Graduates			
	# Employment answers		# Earnings answers		# Employment answers		# Earnings answers	
	M	F	M	F	M	F	M	F
1998	852	1,170	184	261	1,091	1,228	219	221
1999	876	1,109	192	253	1,154	1,197	225	204
2000	909	1,113	206	248	1,258	1,231	242	200
2001	765	1,021	165	218	1,160	1,232	211	202
2002	702	1,028	136	206	1,092	1,181	230	203
2003	826	927	168	195	1,073	1,144	201	197
2004	779	962	156	209	1,038	1,113	213	180
2005	785	865	135	160	1,009	1,283	171	218
2006	694	851	123	165	1,105	1,221	199	161
2007	692	820	108	146	1,045	1,277	175	195
2008	659	715	98	109	1,215	1,235	182	207
All	8,539	10,581	1,671	2,170	12,240	13,342	2,268	2,188

Table 6: LFS unweighted samples sizes in England in 2011/12. Shows number of employment and earnings answers by ages which are the closest match to SLC cohort and gender. The earnings answers is the number of individuals giving positive earnings answers. This question is only available in waves 1 and 5 in the LFS, but is still subject to high rates of non-response. Individuals may appear up to four times in the employment columns but only once in the earnings columns.

Individuals are included if they answer the employment or earnings questions at least once during the four waves for a given tax year (although in no cases does an individual answer the earnings question without answering the employment question). We include all observations here, meaning some individuals will appear up to four times in the employment column. They can only appear once in the earnings column, however, as waves 1 and 5 for any individual cannot occur within 12 months of each other. However, the lower number of earnings answers is not only driven by being asked only in 2 of 5 waves. In addition to this, response rates to the LFS earnings questions are relatively poor, with only around 70% of individuals in employment responding to the earnings questions when asked them (with little difference by gender). In subsequent analysis of earnings distributions we use LFS “piwt” weights which deal with this non-response conditional on being in work. The low sample sizes in the LFS are a concern, so in subsequent analysis we pool across the 1998-2003 cohorts inclusive.

4.1 Summary of differences between the datasets

Table 7 summarises the differences between the LFS and the administrative data. Here we discuss six of these key differences and their likely impact.

First, while the LFS observes graduate status, the GS is the 10% sample of the population of borrowers from the SLC. The GS therefore includes borrowers who did not complete their degree

(as mentioned, this is approximately one tenth of graduates) and excludes graduates who did not borrow. Both of these factors are likely to bias down the GS from the true distribution of graduate earnings, as dropouts are likely to earn less than non dropouts, while non-borrowers are likely to earn more than graduates. This latter conclusion follows as students from wealthier backgrounds are likely to earn more (e.g. Crawford and Vignoles (2014)), although it should be noted that Callender and Jackson (2005, 2008) have suggested poor students are more debt averse and these students are likely to earn a lower return to their HE. Of course, as discussed earlier, self report survey measures of schooling level also suffer from measurement error and hence some graduates in the LFS will also be misclassified. We cannot suggest a direction for this bias.

Second, the GS includes individuals who were domiciled in England at the time they applied to HE. We do not observe this in the LFS, and instead focus on graduates living in households in England at the point of survey. For the GS a major drawback is that amongst those who have moved abroad we do not observe their earnings. The likely scale of this problem is further discussed in Section 3.1.4, but this will bias down estimates. One major cause for concern here was the possibility that individuals from overseas would reside in England for long enough to qualify for loans, get a loan, pay off quickly and move abroad. However, the distribution of earnings amongst those who clear their debts within one year of graduating does not look very different to the baseline, suggesting this is not driving the results more than could already be accounted for from Table 3. For the LFS, English students who moved abroad will not be included at all, while graduates living in England at the point of survey but who studied abroad will be included. Further, the LFS only includes those living ‘households’ which will miss those in the army and some postgraduates living in student accommodation. The sign of each of the biases that arise from population differences is unclear and could go in either direction.

Third, while we observe precise cohort (i.e. year started HE) in the GS data we do not observe this in the LFS, and therefore have to impute cohort from age. This is likely to create biases in the LFS data, as their population is of younger graduates. However, in the GS, we did not find that assigning everybody to cohorts based on age rather than actual observed cohort makes very much difference to the distribution, suggesting the impact of this is likely to be small.

Fourth, the GS includes actual observed annual taxable earnings, for which there is a legal requirement for accurate reporting to HMRC. For the majority of individuals this reporting comes from their employer. Meanwhile, LFS earnings are self-reported with no legal obligation or checks, and are therefore subject to selection issues from non-response and measurement error. Another potential error in the LFS data is that respondents report their earnings over a sample period chosen by themselves, and this is converted into a weekly figure for researchers to use. For our purposes,

	LFS graduates	Golden Sample
Definition of graduates	Those whose highest qualification is at graduate degree level. For majority this is “higher degree” or “first degree”.	Those who borrowed from the SLC. Includes those who borrowed and failed to complete degree. Excludes those who did not borrow.
Population	Graduates living in England at the point of survey who are surveyed and respond. Those with ‘variable’ earnings and those not in households excluded.	10% sample of English-domiciled (on application) borrowers from the SLC. Includes those never in contact with HMRC and those living outside England.
Definition of cohort	Allocated based on age on August 31 in a given year.	Observed year started borrowing.
Earnings	Gross weekly earnings in first and second job combined, multiplied by 52. Weekly earnings are imputed in the survey based on a response period chosen by the individual.	PAYE & SA reported annual labour income. Individuals are legally required to report.
Pensions	Employer contributions usually excluded. Employee contributions usually included.	Employer & employee contributions excluded.
Proxy responses	Included (although this has limited impact on the qualitative conclusions of the paper).	Not applicable
Self-employment	Included, but with no earnings data.	Included.

Table 7: Summary of differences between the LFS graduate and the Golden Sample datasets

we multiply this weekly figure by 52 to get annual earnings for comparison with the HMRC data. This can bias earnings in either direction, but is likely to be worst for those with unsteady work or highly variable pay. The LFS attempts to deal with this by excluding the earnings of those who indicate that they have variable pay - these individuals therefore appear as employed but do not have any earnings information. This is likely to disproportionately exclude low-paid individuals.

Fifth, the GS includes Self-Employed earnings while these individuals are excluded from the LFS earnings data (but not from the LFS altogether). As we see in Table 4, the share of self-employed individuals with low earnings (conditional on having non zero earnings) is higher than for the rest of population, meaning they pull down the distribution compared to the LFS. This implies that the LFS distribution is biased upwards compared to the true distribution of graduate earnings, as self-employed individuals should be included in this. However, Table 4 shows that the overall share of self-employed individuals is relatively low, and we also found that excluding Self-Employed earnings from our positive earnings distribution did not have a dramatic impact.

Finally, in the UK there are two types of pension contribution - employer and employee. Employee contributions are tax-free deductions. Both the GS and the LFS exclude employer pension contributions, but while the GS excludes employee pension contributions, it is likely that individuals will report this in the LFS (although there is some ambiguity depending on the respondent’s interpretation). The associated “bias” of this difference depends on what one is interested in measuring: for the taxpayer returns to HE or the long-run cost of income contingent loans, taxable earnings are what is important; for estimating overall individual returns to HE, pension contributions should be included.

In summary, compared to the true distribution of graduate earnings, the majority of the biases in the GS are negative (including dropouts, missing wealthy graduates, including people who move abroad, excluding pension contributions), while for the LFS they are mostly positive (annualising earnings from sometimes shorter periods, excluding those with variable earnings, excluding self-employed earnings, including pension contributions).

When we consider differences between the SS and LFS non-graduates, many of these differences hold. However three additional problems arise that are discussed in more detail in Section 3.2.1, namely the exclusion from the SS of people who never are in contact with HMRC, the possible inclusion of foreign individuals who are in the country for a short space of time and then leave, and the inclusion of graduates who do not borrow (alongside the exclusion of HE dropouts). We believe the second of these is likely to outweigh the first, while we adjust for the third with the correction in Section 3.2.1. Overall, we think the SS is likely to be biased downwards compared to the “true” distribution of non-graduates, therefore.

5 Comparing earnings distributions

In this section we compare the share of individuals with no and low earnings and the positive earnings distributions in the LFS and the administrative data, all done separately by gender. We first compare the GS with graduates in the LFS, then the Corrected SS with non-graduates in the LFS, then the GS and SS combined with the full LFS sample. Throughout this section the Figures we provide are given for the 2008/09 and 2011/12 tax years, with results for 2009/10, 2010/11 and 2012/13 provided in the Online Appendix C. To deal with small sample size issues in the LFS, we pool across the 1998-2003 cohorts of students (as defined above) in each case. An alternative is to use a model-based approach that pools more years of data and cohorts in the estimation, then predicts earnings for a given cohort in a given year. We document this in the Online Appendix D, though in practice we find this approach does not impact our conclusions.

5.1 LFS and Golden Sample Comparison

Table 8 gives the percentages of graduates with no and low earnings for the LFS and the GS, by gender for each of the five tax years from 2008/09. Individuals with no earnings in the tax data either have no form for that given year or have filed a form with zero earnings. Individuals with no earnings in the LFS are those who have indicated they are not in employment (thus including the unemployed and the economically inactive). Overall, the share of individuals with zero earnings is comparable between the datasets, with the difference generally 1-2 percentage points, most of which can be explained by individuals in the GS moving abroad. However, differences emerge when

this is broken down by gender. In the LFS, the share of men on zero earnings declines with age from 9 to 7%, while the share of women on zero earnings increases with age from 12 to 18% in the LFS. Meanwhile the share of both men and women on zero earnings in the administrative data increases with age from around 12% to around 14%. This discrepancy does not appear to be driven by the inclusion of those who have moved abroad in GS, since this share is too small to explain the differences (see Table 3), and does not differ dramatically by gender.

The Table also gives the share with earnings below £8,000, conditional on working. Here the differences between the LFS and GS are stark; only around 5% of those in employment earn below £8,000 in the LFS, while for the GS it is around 14%. There are also clear gender differences in the survey data, where this fraction for females is around double the equivalent for men, while in the GS the gender differences are minimal, with the exception of only the most recent data.

Year	% Not employed						% Earnings < £8,000 given Earnings > £0					
	Sample size			Share			Sample size			Share		
	All	M	F	All	M	F	All	M	F	All	M	F
<i>LFS</i>												
2008/09	9,234	4,088	5,146	10.4	8.7	12.0	2,249	1,012	1,237	4.6	3.2	6.1
2009/10	9,375	3,952	5,423	11.3	9.3	13.1	2,244	931	1,313	5.3	3.6	7.0
2010/11	9,582	4,057	5,525	11.2	8.0	14.1	2,323	955	1,368	5.4	2.8	7.9
2011/12	10,297	4,315	5,982	13.0	8.9	16.5	2,350	1,004	1,346	4.5	1.7	7.4
2012/13	8,596	3,616	4,980	12.5	6.9	17.5	1,882	807	1,075	5.9	3.6	8.1
<i>Golden Sample</i>												
2008/09	132,401	61,492	70,909	11.4	11.9	10.9	117,332	54,187	63,145	13.8	13.9	13.8
2009/10	132,401	61,492	70,909	13.8	14.0	13.7	114,090	52,886	61,204	13.8	14.2	13.4
2010/11	132,401	61,492	70,909	13.4	13.6	13.2	114,669	53,111	61,558	13.5	13.3	13.7
2011/12	132,401	61,492	70,909	13.5	13.4	13.5	114,578	53,268	61,310	13.3	12.7	13.7
2012/13	132,401	61,492	70,909	14.6	14.4	14.7	113,111	52,659	60,452	13.9	12.5	15.2

Table 8: LFS and Golden Sample: graduates not employed and with low earnings overall and by gender for 2008/09 through 2012/13. The 1998-2003 cohort are pooled for each year. LFS population weights are applied (the “pwt” weight for the unemployed share, while the “piwt” weight for the earnings share).

We know from Table 8 that the differences in the shares with low earnings between the datasets are large. In Figure 1 we consider the earnings distributions for the LFS and GS, conditional on earnings being greater than £8,000, to consider the possibility that earnings might be more comparable above this point. However, we see that even here considerable earnings differences exist, with earnings generally being higher in the LFS right through the distribution until you get to the high percentiles. This pattern is more clear for men than for women and is more true in 2008/09 than in 2011/12. It is reflected in the difference in conditional means (given to the right of each panel in the Figure), which are higher in the LFS than in the GS.

Returning to the full earnings distribution in Figure 2, we observe considerably higher earnings in the LFS than the GS in the lower parts of the distribution. The precise numbers are given in the Online Appendix C. In 2008/09, at the 10th percentile, earnings are almost 3 times higher

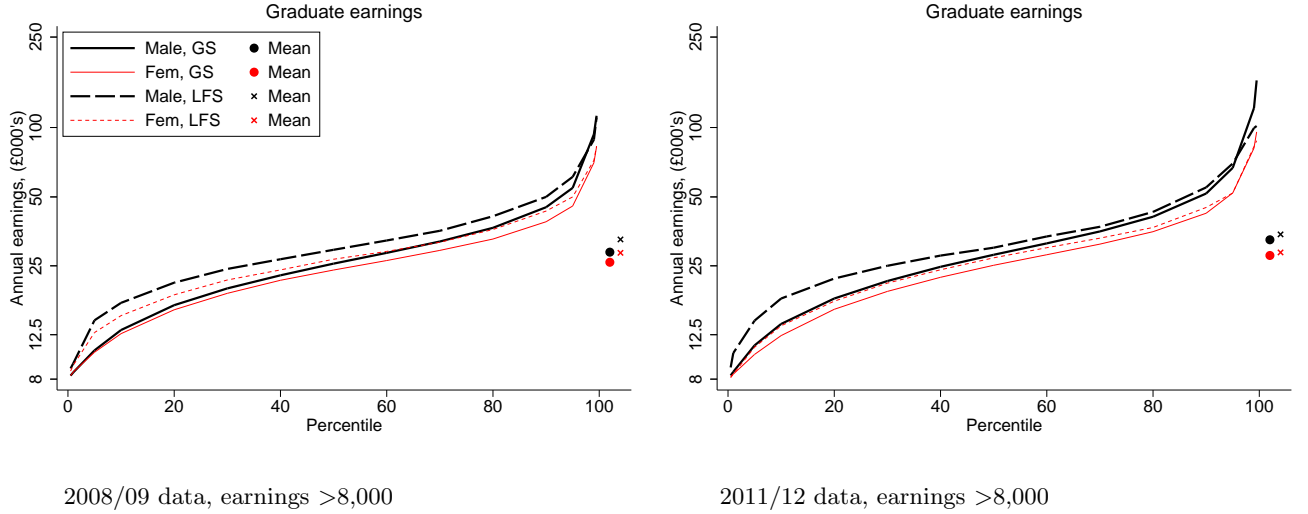


Figure 1: Graduate earnings, 1998-2003 cohorts pooled. Non-parametric estimates of the LFS graduate & GS earnings distributions, for earnings > £8,000. Left hand y-axis shows annual earnings on a log scale and right hand axis shows absolute % difference between the LFS and GS. Conditional means are provided to the right of each picture (horizontal jitter is included to improve clarity). LFS earnings are weighted using population weights.

for men in the LFS (£15,900) than in the GS (£5,600) and twice as high for women (£12,100 and £5,900) respectively. The gap declines in percentage terms at higher levels of earnings, but persists through the distribution up to the top tail. This is true for both genders in each of the five years we investigate, with earnings only ever higher in the GS at or above the 99th percentile of the distribution. Consequently mean earnings (conditional on working) are always higher in the LFS - by around 20% for men and slightly less for women. Indeed there is little difference in earnings between male and female graduates in their mid-late twenties and early thirties in the GS, while the differences in the LFS are larger.

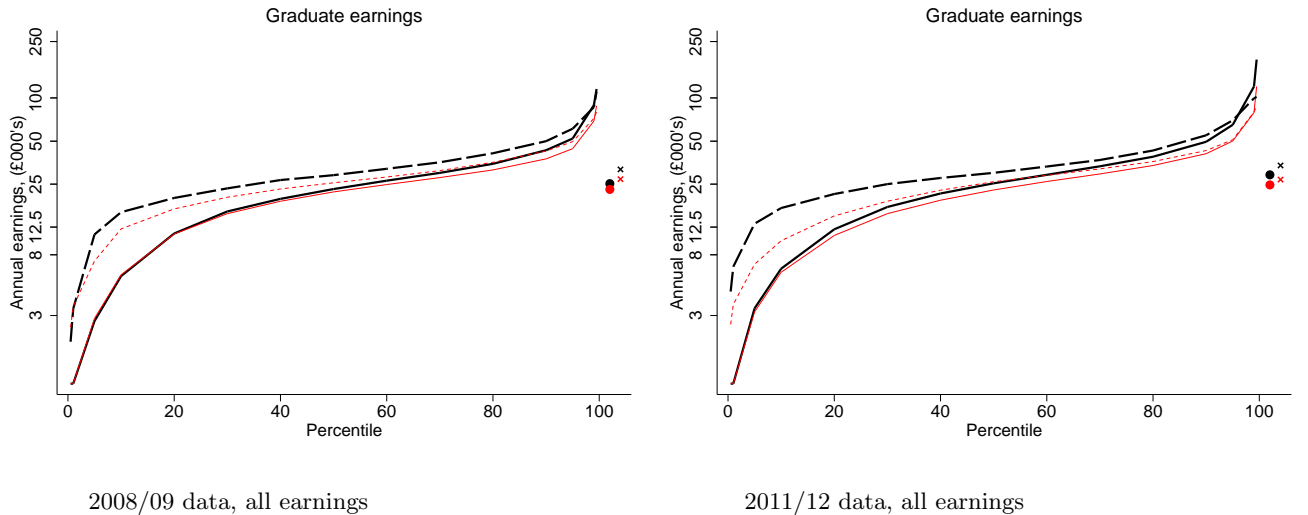


Figure 2: Mimics Figure 1 but with the full earnings distribution, not including zeros. See Figure 1 for legend. Precise numbers given in the Online Appendix C, along with other years of data.

The discrepancies between the GS and the LFS are striking and we consider a few potential explanations. First, lower earners may be less likely to respond in the LFS. Lower income individuals are more likely to be doing shift work, have varied hours and be more geographically mobile, which may make it more difficult to get hold of them longitudinally to complete the survey. Further, the low paid may be less inclined to complete the survey. For this to be the driving source of differences, this response issue must be gendered, with low-earning men considerably less likely to respond than low-earning women.

Second, LFS reported earnings are subject to measurement error, caused by annualising earnings from sometimes shorter periods. Around 60% report earnings for periods of less than a year, and those individuals typically have lower earnings than those who report earnings over a one year period. Individuals with very low earnings are likely to spend periods of the year out of work and this is ignored in such calculations. This explanation is supported by the fact that the discrepancy between the GS and the LFS is larger in 2008/09 than in 2011/12, a time when the labour market was more turbulent, but would be more convincing were the share of those on zero earnings greater in the LFS than in the administrative data, which is not the case. A more convincing explanation is that the LFS excludes the earnings of individuals with variable pay. Our analysis here essentially assumes that the data are randomly missing earnings information from the set of individuals in employment. In practice this is unlikely, and this is instead likely to underestimate the share of individuals with low earnings.

Third, the LFS excludes self employed earnings. Referring back to Table 4 we see that around 45% of the fully self-employed earn below £8,000. However, since this only accounts for less than 4% of individuals, excluding these individuals from the tax data would only reduce the share earning between £0 and £8,000 by around 1.5 percentage points, a small fraction of the overall difference.

Fourth, employee pension contributions are excluded from the GS. This could plausibly explain a large share of the differences in earnings between the datasets above the median. According to the ONS figures, average employee contributions are 6% of earned income, and although our own calculations from the British Household Panel Survey suggest these are much smaller for 20-30 year olds, they increase considerably with earnings - we estimate they explain around half of the differences above the median for men and more for women. However, this is highly unlikely to explain the differences at the bottom of the distribution.

A final explanation is that the differences are instead driven by under-reporting of earnings in the GS. There may be earnings that people simply do not report to the tax office, such as earnings for cash in hand overtime work, to avoid paying tax or to receive in-work benefits. This is likely to be important when looking at the low paid. This would result in the proportion of

individuals reporting very low earnings being higher in the official data than in the LFS, with the latter collecting a more realistic assessment of total earned income. However, this issue may also be relevant when considering gender differences; from the unconditional figures, it is clear that the LFS shows a stronger gender gap than the GS. Potentially this would imply that males are more likely to have second jobs and additional sources of income that they include in their reporting to the LFS but that they do not declare to the tax office. Our conclusions for the LFS are unchanged by excluding earnings from second jobs but it is possible that individuals provide aggregate estimates of their earnings from all jobs when responding to the income questions in the LFS and hence this could still be an explanation.

In summary, there are a number of potential reasons why earnings are lower in the GS than the LFS and we have no definitive explanation, particularly for the share of low earners. It is however, important from a policy perspective to observe that the official tax record is the relevant one in terms of both tax contributions and repayment of student loans.

5.2 LFS non-graduate and Administrative data non-HE comparison

We now turn our attention to non-graduates. In the LFS these individuals are those who have not achieved a higher degree or equivalent, while for the administrative data we use the Corrected SS, as described in Section 3.2.1. Table 9 compares the share of individuals not employed and, conditional on working, the share reporting earnings between £0 and £8,000. As for graduates, while the overall share of those not in employment is very similar across the two surveys, there are considerable gender differences in the LFS which do not exist in the administrative data. As previously described, we were particularly concerned about the share of low earnings in the SS due issues with the ‘never filers’, i.e. those who never file with HMRC, and with immigration. For these factors to be driving the results, ‘never filers’ would have to be predominantly women, while immigrants would have to be predominantly men. While this is plausible, these patterns are similar for the GS where these issues do not apply.

Figure 3 shows the distribution of earnings for those with earnings above £8,000 and for the entire distribution combined. Above £8,000, the distributions are quite similar, with the LFS reporting higher earnings at the lower end of the distribution and lower earnings at the higher end. For the full distribution, we have a similar story as for the graduate comparison, with the LFS reporting considerably higher earnings at the low end of the distribution. There is again more gender disparity in the survey data than in the tax data, although there is clearly a much greater gender difference in the nonHE sample than in the GS.

Some of the potential explanations rehearsed in respect of differences between the GS and the LFS for graduates’ earnings also apply to non-graduates’ earnings. In addition to this, the

Year	% Not employed						% Earnings < £8,000 given Earnings > £0					
	Sample size			Share			Sample size			Share		
	All	M	F	All	M	F	All	M	F	All	M	F
<i>LFS</i>												
2008/09	16,326	7,759	8,567	27.0	17.2	37.3	2,942	1,484	1,458	11.7	4.2	20.9
2009/10	14,920	7,021	7,899	30.1	21.0	40.0	2,594	1,326	1,268	13.2	4.8	24.5
2010/11	14,382	6,925	7,457	28.2	17.5	40.1	2,430	1,261	1,169	13.0	4.1	25.5
2011/12	14,041	6,828	7,213	28.3	18.1	39.8	2,481	1,278	1,203	13.9	4.3	28.8
2012/13	10,495	5,082	5,413	27.6	16.3	40.1	1,778	946	832	15.4	5.4	29.9
<i>HMRC non-HE</i>												
2008/09	243,099	132,522	110,577	29.1	28.5	30.0	172,245	94,816	77,429	37.4	33.7	42.0
2009/10	243,099	132,522	110,577	31.4	30.1	32.9	166,805	92,588	74,217	37.6	34.5	41.5
2010/11	243,099	132,522	110,577	29.9	28.8	31.2	170,451	94,392	76,059	38.0	34.7	42.0
2011/12	243,099	132,522	110,577	29.0	27.7	30.6	172,581	95,870	76,711	38.9	34.7	44.3
2012/13	243,099	132,522	110,577	28.8	27.6	30.2	173,102	95,966	77,136	38.7	33.9	44.7

Table 9: LFS and Corrected Silver Sample: non-graduates not employed and with low earnings overall and by gender for 2008/09 through 2012/13. The 1998-2003 cohort are pooled for each year. Note the share not employed in the Corrected Silver Sample is equal to the share no employed in the Silver Sample as the econometric correction only corrects the positive earnings distribution.

immigration issue also affects the positive earnings distribution if overseas workers work for a fraction of the year then leave the country. However, this is unlikely to be the whole story, and the very similar pattern for graduates suggests similar factors might be at work.

5.3 Graduate and non-graduate combined comparison

Finally, we compare the LFS with the combined GS and SS in Figure 4. We use the SS rather than the corrected SS distribution as we are looking at the whole distribution together, so it does not matter if some graduates are misclassified as being in the SS. Conditional on earnings being above £8,000, the LFS and the combined administrative data earnings distributions are somewhat more similar. When considering the full distribution, unsurprisingly the familiar patterns again emerge, with a much higher share of low earners and much less gender disparity in the administrative data.

Given these discrepancies, we have also (with help from colleagues at the IFS) undertaken a comparison of the share earning below £8,000, conditional on working, using the Family Resources Survey (FRS), another commonly used data source with information on workers' earnings. The FRS suggests around 17% of men and 27% of women in their late 20's were earning between £0 and £8,000 in 2011/12 (in 2012 prices). These numbers are both higher than the LFS, considerably so for men. The female share is very close to the administrative data figure for women, but still considerably lower for men. Our reasoning above would suggest that under-reporting of earnings is likely to be a significant cause of this but again we stress that the official tax record is the relevant data for many important policy purposes, including repayment of student loans.

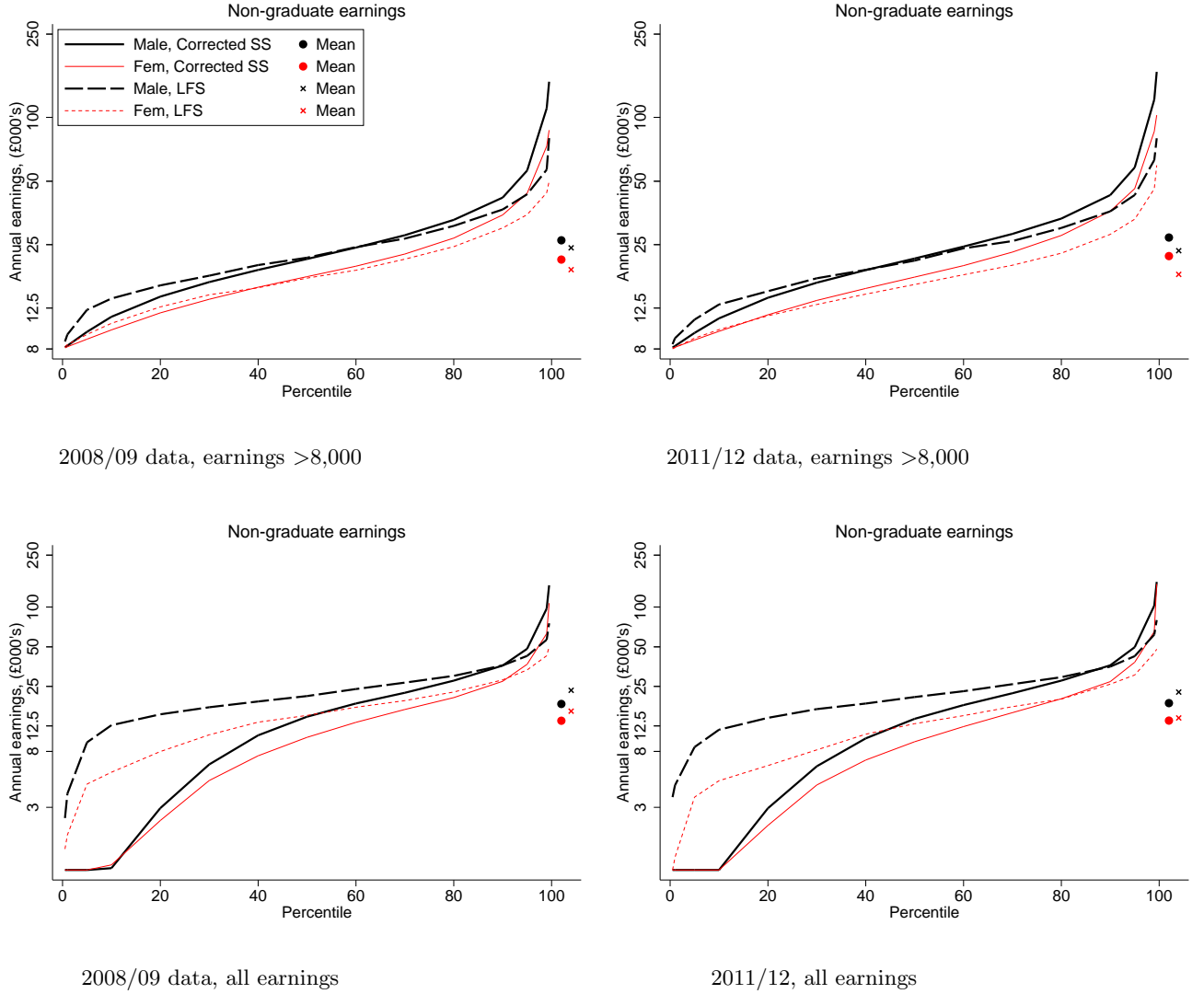


Figure 3: Non-graduate earnings, 1999 cohort. Mimics Figure 1 and 2. Precise numbers given in the Online Appendix.

6 Applied comparisons

Thus far we have shown comparisons of the cross sectional distributions of the administrative and survey data. In this section we turn our attention to the implications of these findings for measuring the gender wage gap, the graduate premium and earnings inequality. In our Online Appendix E we also show graduate and non-graduate earnings growth in the five years following the recession. Each of these measures is of crucial importance for policy.

6.1 The gender wage gap

The gender wage gap is of considerable policy interest as large differences in pay by gender are known to have existed for a long time. We explore this in Figure 5, which shows the gap at various points in the positive earnings distribution, for the LFS and administrative data. We do this for

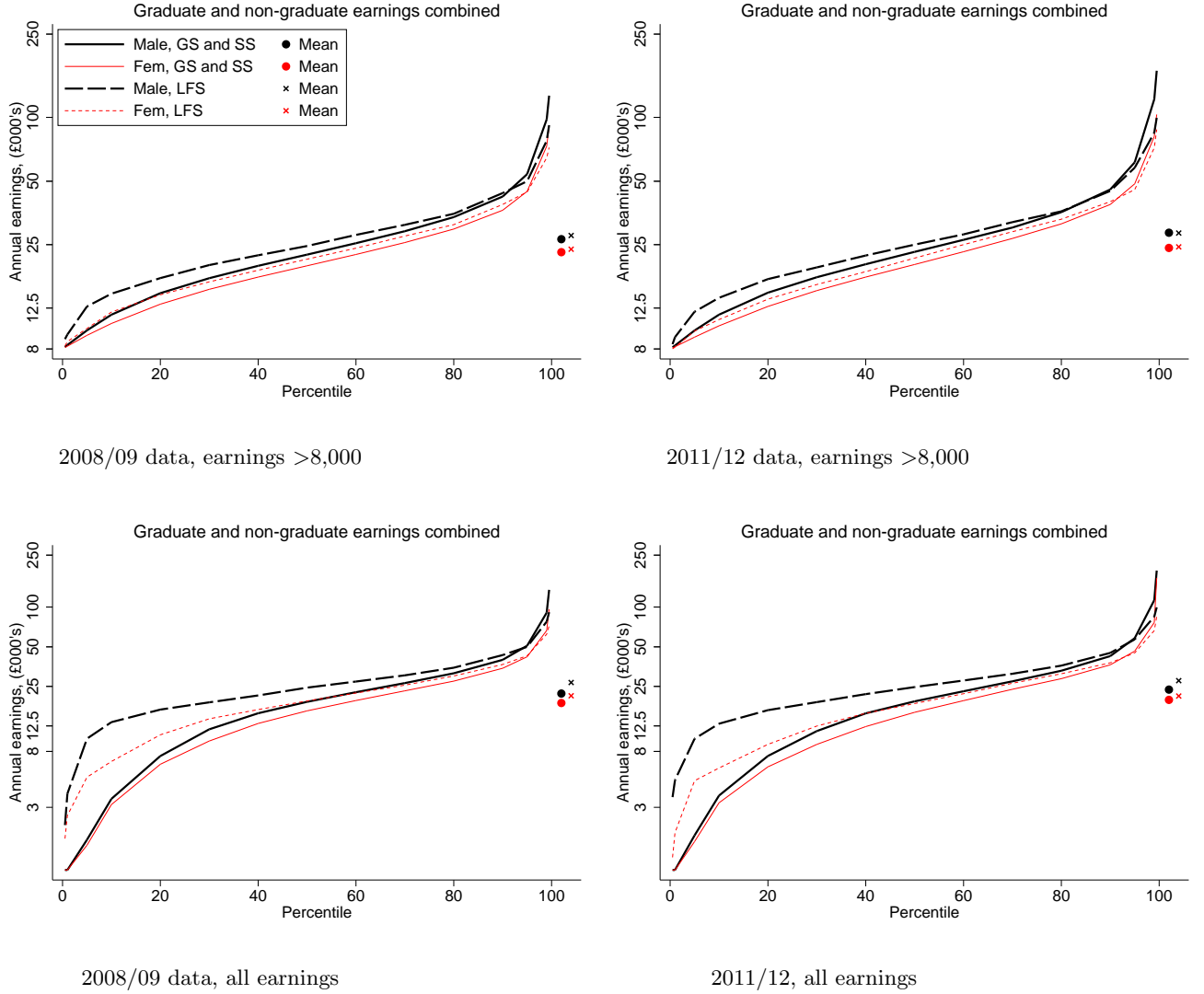


Figure 4: Graduates and non-graduates combined, 1999 cohort. Mimics Figure 1 combined with Figure 2. Precise numbers given in the Online Appendix.

both graduates and for non-graduates, showing ratios through the distribution for 2008/09 and 2011/12 for the 1998-2003 cohorts combined. Ratios for 2009/10, 2010/11 and 2012/13 are shown in the Online Appendix F.

We find that the pay gap between men and women is larger for non-graduates than for graduates, suggesting HE plays a role in alleviating gender differences in earnings. This result is more pronounced in 2011/12 than 2008/09, which could be driven by age effects. Above the 40th percentile, we find that the LFS and administrative data ratios are actually very similar, at around 1.1-1.2 for graduates and 1.4-1.5 for non-graduates. At the bottom of the distribution, however, the data sources give very different results, with much less gender difference in the administrative data. These differences have further impacts on the mean male:female earnings ratio, with the survey data suggesting a greater wage gap than the administrative data.

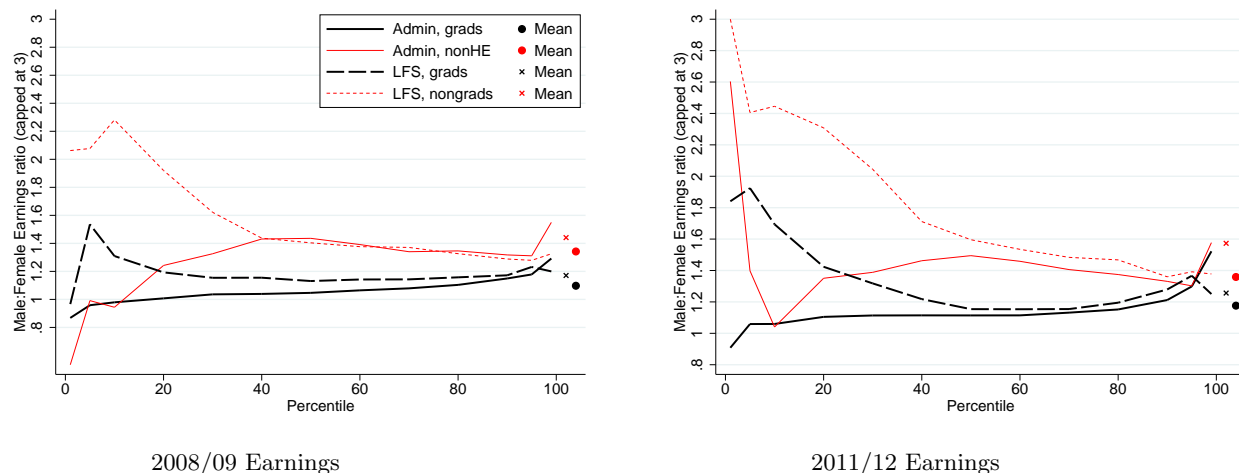


Figure 5: Ratio of male vs. female earnings at different points in the distribution in the administrative data & in the LFS in 2008/09 & 2011/12 for the 1998-2003 cohorts combined, by graduate status. See Online Appendix for raw numbers behind these figures and for other years.

6.2 Graduate vs. non-graduate earnings

The ratio of graduate to non-graduate earnings is important for those considering the value of getting a degree (we refer to this as the graduate premium, though the differences here are descriptive rather than causal). In Figure 6 we show the ratio of graduate to non-graduate earnings at various points in the positive earnings distribution for 2008/09 and 2011/12 (other years are provided in the Online Appendix F), for the 1998-2003 cohorts combined.

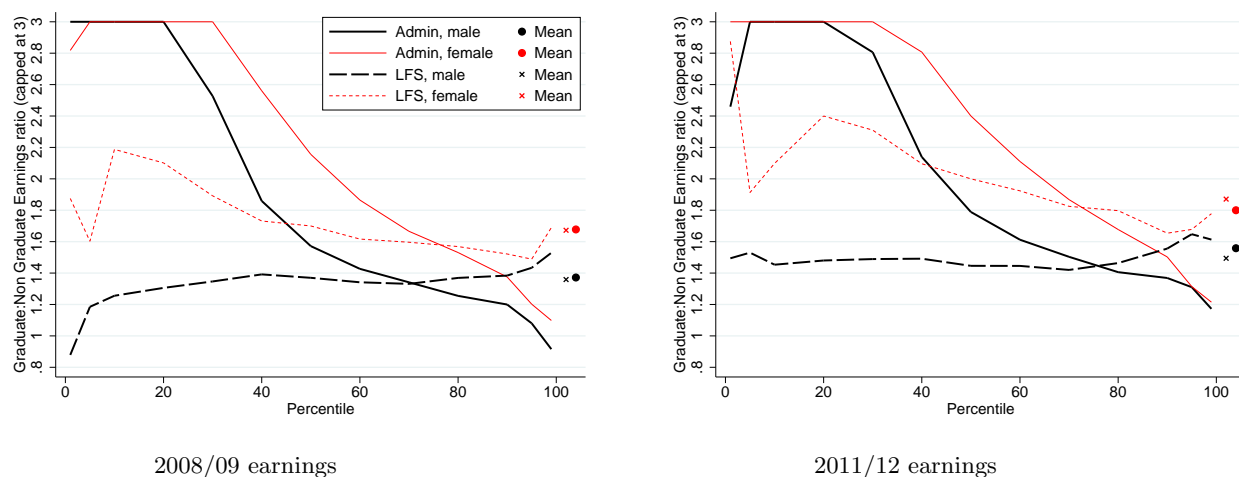


Figure 6: Ratio of graduate vs. non-graduate earnings at different points in the distribution in the administrative data & in the LFS in 2008/09 & 2011/12 for the 1998-2003 cohorts combined, by gender. See Online Appendix for raw numbers behind these figures and for other years. The ratio is capped at 3.

The graduate earnings premium is larger in the administrative data than in the LFS at the lower percentiles of the earnings distribution, though the premia in the two data sets converge further up the distribution. On average across the two datasets the graduate wage premium is around 1.7 (1.4) for women (men) in 2008/09 and 1.8 (1.5) in 2011/12. Hence the graduate premium is

larger for both women and men in 2011/12 than in 2008/09, suggesting more growth in graduate earnings for individuals in their late twenties and early thirties. At the very highest percentiles, the administrative data shows lower graduate:non-graduate ratios. This may partially reflect the inadequacies of the non-HE sample, which is the result of an econometric correction to the SS to allow for the fact that the SS includes graduates who do not borrow. It is possible that the correction we use is particularly weak at the higher end of the distribution due to the increased likelihood of the presence of graduates in that part of the distribution.

6.3 Earnings inequality

Finally, we consider earnings inequality in our different data sources. In each case, we sort n earnings as $Y_{[1]}, Y_{[2]}, \dots, Y_{[n]}$. The Lorenz (1905) curve plots $L_n(s)$, the cumulative share of earnings against the population fraction $s \in [0, 1]$,

$$L_n(s) = \frac{\sum_{j=1}^{\lfloor ns \rfloor} Y_{[j]}}{\sum_{j=1}^n Y_{[j]}},$$

where $\lfloor x \rfloor$ generically denotes the integer part of x . The Gini coefficient $G_n = 2 \int_0^1 \{s - L_n(s)\} ds$ summarises the curve as twice the area between the 45° line and the curve (alternative measures include the Atkinson (1970) index). In Table 10 we report Gini coefficients for 2008/09 through 2012/13, for each year pooling across the 1998-2003 cohorts.

Year	Graduates				Non-Graduates			
	Men		Women		Men		Women	
	GS	LFS	GS	LFS	CSS	LFS	CSS	LFS
2008/09	0.360	0.259	0.332	0.263	0.505	0.246	0.509	0.304
2009/10	0.372	0.252	0.337	0.268	0.510	0.240	0.515	0.319
2010/11	0.380	0.268	0.347	0.272	0.525	0.255	0.520	0.355
2011/12	0.388	0.262	0.358	0.283	0.523	0.255	0.540	0.340
2012/13	0.395	0.283	0.375	0.287	0.528	0.251	0.540	0.370

Table 10: Gini coefficients for the administrative data and LFS positive earnings distributions, split by gender and graduate status for 2008/09 - 2012/13. Each observation includes the 1998-2003 cohorts. The ‘GS’ (Golden Sample) and ‘CSS’ (Corrected Silver Sample) show the administrative data. See Tables 8 and 9 for sample sizes.

The Table shows that inequality generally rises between 2008/09 and 2012/13, probably due to a combination of the financial crisis and age effects as we hold cohorts fixed. There is considerably less earnings inequality in the LFS than in the administrative data, which is unsurprising given the dramatically higher earnings in the LFS at the bottom of the distribution. Earnings inequality is much greater for non-graduates than for graduates in the administrative data. This pattern is less clear for males in the LFS.

7 Conclusion

This paper compares earnings distributions from administrative tax records with LFS survey data. Understanding the strengths and weaknesses of these types of datasets and their use in the study of graduates' earnings is important as the use of administrative data by policy makers becomes more prevalent in the UK and worldwide. Broadly we find that the administrative data show lower mean earnings for both graduates and non-graduates and in particular suggest a far greater proportion of individuals with earnings below £8,000 than does the LFS.

We explored a number of possible reasons for this difference in the distributions, including under-reporting of earnings in the administrative data and response bias and measurement error in the survey data. Although the LFS is likely to suffer from biases, the administrative data also suffer from biases that particularly affect the lower parts of the distribution. In addition, there are some specific issues when using the administrative data to analyse the earnings of graduates. Specifically, the GS records the earnings of individuals who study in England but then move abroad as zero, includes individuals who drop out of their degree without graduating, and does not include graduates who did not borrow. All three of these factors could also bias downwards estimates of graduate earnings. However, the fact that we get a very similar pattern of differences when comparing the LFS with the administrative data for non-graduates suggests that the unique selection issues of the GS are not the main drivers of the differences at the bottom end of the earnings distribution. More likely explanations include sample selection resulting in low response in the LFS from low earners - in particular the exclusion of those with variable pay - that is not sufficiently captured in the population weights, systematic over-reporting of earnings in the LFS by low earners, or significant under-reporting of income from employment by lower earners in the administrative data. There is some evidence for the former problems, as estimates of graduates' earnings using the FRS are far more closely aligned with the administrative data, particularly for females, though there is still a higher proportion of males in the administrative data with very low earnings. The under-reporting issue in the administrative data could perhaps be due to people doing casual unreported work or due to shifting their earnings into other forms of income, to avoid paying tax or in order to qualify for working tax credits. The greater share of self-employed people with no or low earnings seems to support this.

These differences between the two data sets have several important implications for our empirical findings. The LFS data suggest less earnings inequality, particularly for non-graduate men and a considerably larger gender gap. We also find that the LFS data paint a less favourable picture of the economic advantage of HE, as it exhibits a smaller graduate to non-graduate earnings ratio. We also show in the Online Appendix E that the administrative data display a smaller negative earnings

shock for graduates in the years following the Great Recession - in both datasets, the decline in real earnings after the recession is large for graduates but the decline for non-graduates is larger in the administrative data. Hence overall, the differences between the data sets are substantively important for policy research.

In summary, the new administrative dataset has great potential for research, and may result in different conclusions about important labour market issues. However, we also raise issues about the reliability of the administrative data at the lower end of the earnings distribution which merit further debate and study. Overall we might think that the administrative data are likely to be more reliable than survey data at the upper end of the distribution due to its more comprehensive coverage and the legal obligation for accurate reporting. Further, it is of course the official earnings record on which calculations of tax take and graduate loan repayments are based, and therefore of great practical importance. Improving our understanding about whether the official earnings data are significantly under-reporting the earnings of individuals at the lower end of the earnings distribution is also a pressing issue, not just to ensure that we have accurate information on individuals' earnings for tax purposes but also because of the current policy importance of earnings inequality and its apparent sensitivity to the data source being used to measure it.

References

- Abowd, J. and M. Stinson (2013). Estimating measurement error in annual job earnings: A comparison of survey and administrative data. *Review of Economics and Statistics*.
- Atkinson, A., J. Micklewright, and N. Stern (1981). A comparison of the FES and NES 1971 - 1977: Part I. Characteristics of the sample. Social Science Research Council Programme on Taxation, Incentives and the Distribution of Income Working Paper, (27).
- Atkinson, A., J. Micklewright, and N. Stern (1982). A comparison of the FES and NES 1971 - 1977: Part II. Hours and earnings. Social Science Research Council Programme on Taxation, Incentives and the Distribution of Income Working Paper, (32).
- Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory* 2, 244–263.
- Bhuller, M., M. Mogstad, and K. G. Salvanes (2017). Life cycle earnings, education premiums and internal rates of return. *Journal of Labor Economics*. Forthcoming.
- Black, S., P. J. Devereux, and K. G. Salvanes (2005). The more the merrier? The effect of family size and birth order on children's education. *Quarterly Journal of Economics* 120.2, 669–700.
- Bound, J., C. Brown, and N. Mathiowetz (2001a). Measurement error in survey data. *Handbook of econometrics*.
- Bound, J., C. Brown, and N. Mathiowetz (2001b). Measurement error in survey data. In J. J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics, Volume 5*, pp. 3705–3843.

- Britton, J., N. Shephard, and A. Vignoles (2015). Comparing sample survey measures of English earnings of graduates with administrative data during the great recession. IFS Working Paper W15/28.
- Callender, C. and J. Jackson (2005). Does fear of debt deter students from higher education? *Journal of Social Policy* 34, 509–540.
- Callender, C. and J. Jackson (2008). Does fear of debt constrain choice of university and subject of study? *Studies in Higher Education* 33, 405–429.
- Card, D., R. Chetty, M. Feldstein, and E. Saez (2010). Expanding access to administrative data for research in the United States. Unpublished paper: Dept. of Economics, Harvard University.
- Carneiro, P., T. L. Garcia, K. G. Salvanes, and E. Tominey (2013). Intergenerational mobility and the timing of parental income. CES Ifo Conference on Economics of Education, CES Ifo Area Conferences. CES Ifo; September 67, 2013.
- Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review* 104, 2593–2632.
- Chetty, R., J. N. Friedman, and J. E. Rockoff (2014b). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review* 104, 2633–2679.
- Chevalier, A. (2007). Education, occupation and career expectations: determinants of the gender pay gap for UK graduates. *Oxford Bulletin of Economics and Statistics* 69(6), 819–42.
- Crawford, C. and A. Vignoles (2014). Heterogeneity in graduate earnings by socio-economic background. Unpublished paper: Institute for Fiscal Studies.
- Cribb, J. and R. Joyce (2015). Chapter 2: Earnings since the recession. IFS Green Budget 2015.
- Cunha, F. and J. Heckman (2016). Decomposing trends in inequality in earnings into forecastable and uncertain components. *Journal of Labor Economics* 34(S2), S31–S65.
- Devereux, P. J. and R. A. Hart (2010). Forced to be rich? Returns to compulsory schooling in Britain. *Economic Journal* 120, 1345–1364.
- Duncan, G. J. and D. H. Hill (1985). An investigation of the extent and consequences of measurement error in labor – economic survey data. *Journal of Labor Economics*, 508–532.
- Figlio, D. N., K. Karbownik, and K. G. Salvanes (2015). Education research and administrative data. Nation Bureau of Economic Research, No. w21592.
- Guvonen, F., G. Kaplan, and J. Song (2014). The glass ceiling and the paper floor: Gender differences among top earners, 1981-2012. Unpublished paper, Princeton University.
- Her Majesty’s Revenue and Customs (2014). Measuring tax gaps 2014 edition: tax gap estimates for 2012-13. Issued by Corporate Communications, HMRC.
- Koijen, R., S. Van Nieuwerburgh, and R. Vestman (2015). Judging the quality of survey data by comparison with ‘truth’ as measured by administrative records: Evidence from SWEDEN. *Chapter in NBER Book Improving the Measurement of Consumer Expenditures*, Christopher Carroll, Thomas Crossley, John Sabelhaus, eds., 2015.
- Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association* 9, 209–219.

- Machin, S. and P. Puhani (2003). Subject of degree and the gender wage differential: evidence from the UK and Germany. *Economic Letters* 79(3), 393–400.
- Micklewright, J. and S. V. Schnepf (2010). How reliable are income data collected with a single question? *Journal of the Royal Statistical Society, Series A* 173, 409–429.
- Moore, J. C., L. L. Stinson, and E. J. Welniak (2000). Income measurement error in surveys: A review. *Journal of Official Statistics* 16, 331–362.
- Naylor, R., J. Smith, and S. Telhaj (2016). Graduate returns, degree class premia and higher education expansion in the UK. *Oxford Economic Papers* 68, 525–545.
- Pope, T. and B. Roantree (2014). A survey of the UK tax system. Institute for Fiscal Studies.
- Rodgers, W. L., C. Brown, and G. J. Duncan (1993). Errors in survey reports of earnings, hours worked, and hourly wages. *Journal of the American Statistical Association* 88, 1208–1218.
- Savage, M. and R. Burrows (2009). Some further reflections on the coming crisis of empirical sociology. *Sociology* 43, 762–772.
- Skinner, C., N. Stuttard, G. B. Durrant, and J. Jenkins (2002). The measurement of low pay in the UK Labour Force Survey. *Oxford Bulletin of Economics and Statistics* 64, 653–676.
- Walker, I. and Y. Zhu (2011). Differences by degree: evidence of the net financial rates of return to undergraduate study for England and Wales. *Economics of Education Review* 30, 1177–1186.
- Webber, R. (2009). Response to 'The coming crisis of empirical sociology: an outline of the research potential of administrative and transactional data'. *Sociology* 43, 169–178.
- Wilkinson, D. (1998). Towards reconciliation of NES and LFS earnings data. *Labour Market Trends*.

Acknowledgement

This paper is a revised version of “Comparing sample survey measures of English earnings of graduates with administrative data during the Great Recession.” Here we exclude earnings dynamics. Many civil servants and policy makers have helped us gain access to the data which is the core of this paper. We thank in particular Daniele Bega, Dave Cartwright, Nick Hillman, Tim Leunig and David Willetts for their invaluable contributions. We also thank A.B. Atkinson, Raj Chetty, Jonathan Cribb, Mark Gittos, Chuka Ilochi and Jonathan Waller for their comments previous drafts, and our advisory group, Alison Alden, Nick Barr, Danny Dorling, Josh Hillman, Robin Naylor, Kate Purcell and Ian Walker. We also thank our three anonymous referees for excellent comments. We solely are responsible for any errors. For financial support we are grateful to the Nuffield Foundation for original funding, while Jack Britton is also thankful to the British Academy. The views expressed are those of the authors and not necessarily those of the Foundation or the British Academy.

HM Revenue & Customs (HMRC) and Student Loans Company (SLC) have agreed that the figures and descriptions of results in the attached document may be published. This does not imply HMRC's or SLC's acceptance of the validity of the methods used to obtain these figures, or of any analysis of the results. Copyright of the statistical results may not be assigned. This work contains statistical data from HMRC which is Crown Copyright and statistical data from SLC which is protected by Copyright, the ownership of which is retained by SLC. The research datasets used may not exactly reproduce HMRC or SLC aggregates. The use of HMRC or SLC statistical data in this work does not imply the endorsement of either HMRC or SLC in relation to the interpretation or analysis of the information.